Supporting Material

Incorporating chromatin accessibility data into sequence to expression modeling

Pei-Chen Peng[1], **Md. Abul Hassan Samee**[1] and **Saurabh Sinha**[1,2,§]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
[2]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[§]Corresponding author

E-mail addresses:
PP: ppeng5@illinois.edu
MAHS: samee1@illinois.edu
SS: sinhas@illinois.edu

TEXT S1 Supplementary Methods

Model training

Three different goodness-of-fit functions were used at various stages of optimization, to compare between real and predicted expressions of enhancer sequences: average correlation coefficient (Avg. CC), root mean square error (RMSE), and weighted Pattern Generating Potential (wPGP, taken from (1) and described in the following section). To avoid being trapped in local optima parameter optimizations were done in multiple runs while alternating between Avg. CC and RMSE as the objective functions. The optimization starts with a set of default parameters and Avg. CC as the objective function. Upon convergence, the resulting set of parameters is used to initiate optimization with RMSE as the objective function, which is run to convergence. This procedure of optimizations alternating between Avg. CC and RMSE as objective functions is repeated twice, and the resulting set of parameters initiates the final optimization step that uses wPGP as the objective function. Each optimization is done by alternating between the Nelder-Mead simplex method and the quasi-Newton method, as in (2).

Evaluation of model predictions using wPGP (weighted pattern generating potentials)

Given the predicted and real expression profiles, the wPGP score is defined as follows:

$$\text{wPGP} = 0.5 + 0.5\times \text{(reward-penality)},$$

where $\text{reward} = \frac{\sum_i r_i \times \min(r_i, p_i)}{\sum_i r_i \times r_i}$, and $\text{penality} = \frac{\sum_i(\max_r - r_i) \times (p_i - r_i) \times \text{I}(p_i > r_i)}{\sum_i(\max_r - r_i) \times \sum_i(\max_r - r_i)}$. Here, p_i and r_i are the predicted and the real expression in bin i, respectively, $\max_r$ is the maximum level of real gene expression, and I(B) is a binary variable indicating the truth of condition "B". The wPGP score ranges from 0 to 1, with higher scores indicating better matches between the predicted and the endogenous expression. The wPGP score was used as the objective function during parameter training, as well as for assessing if one model fits the data better than another.

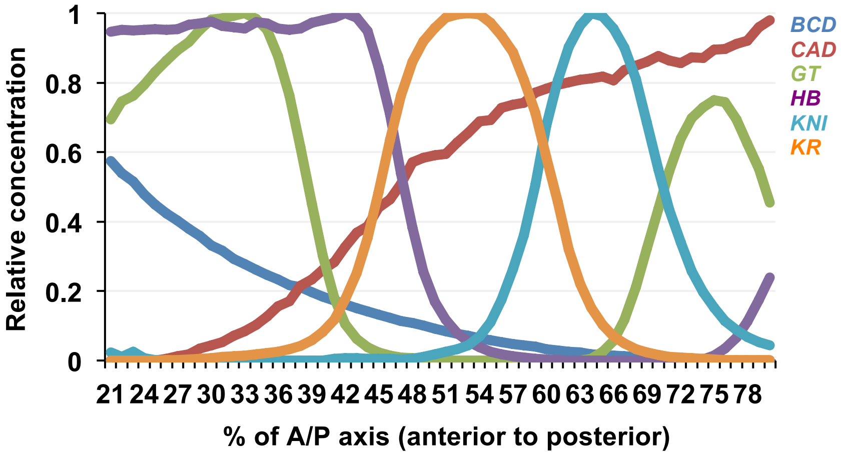
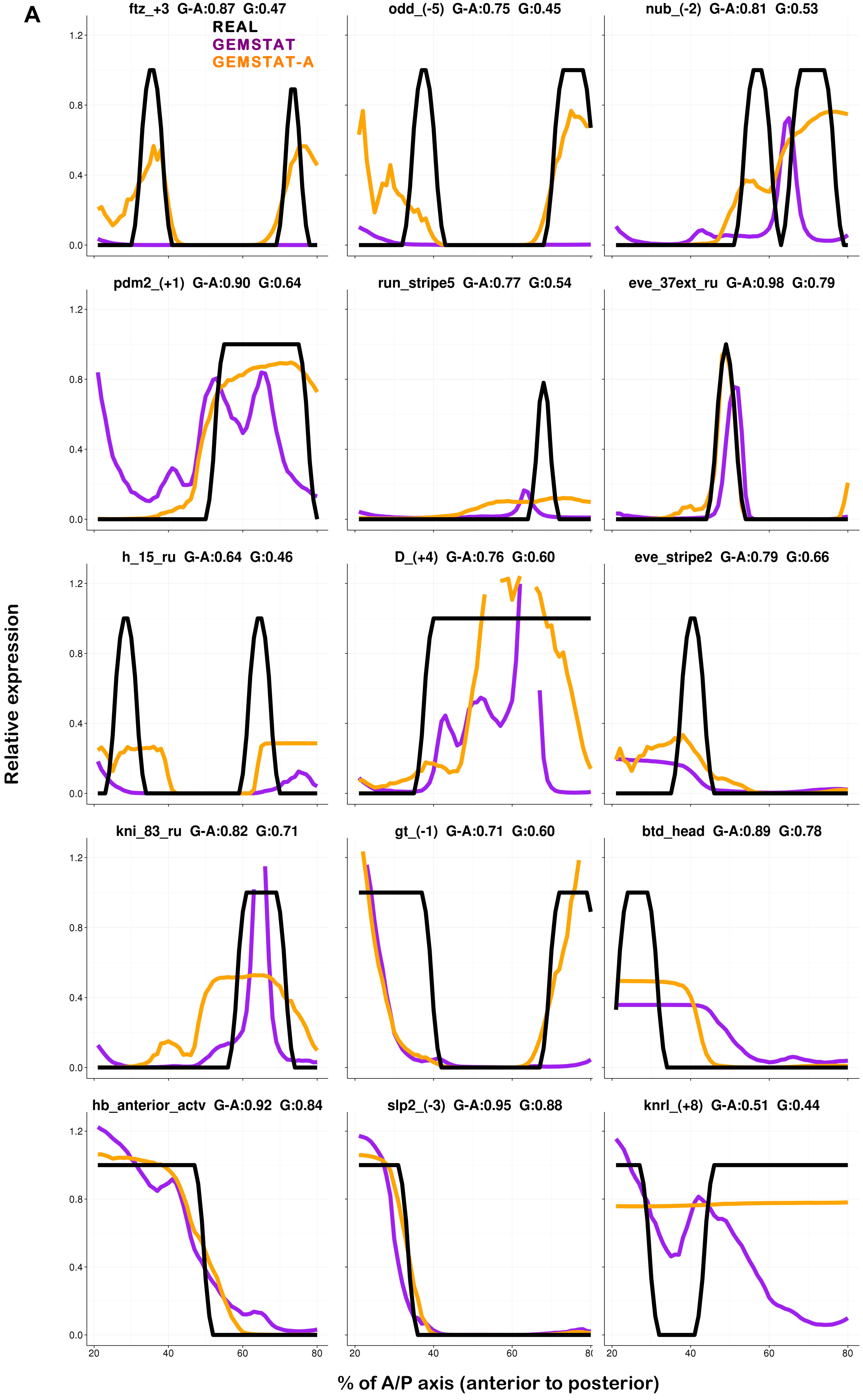
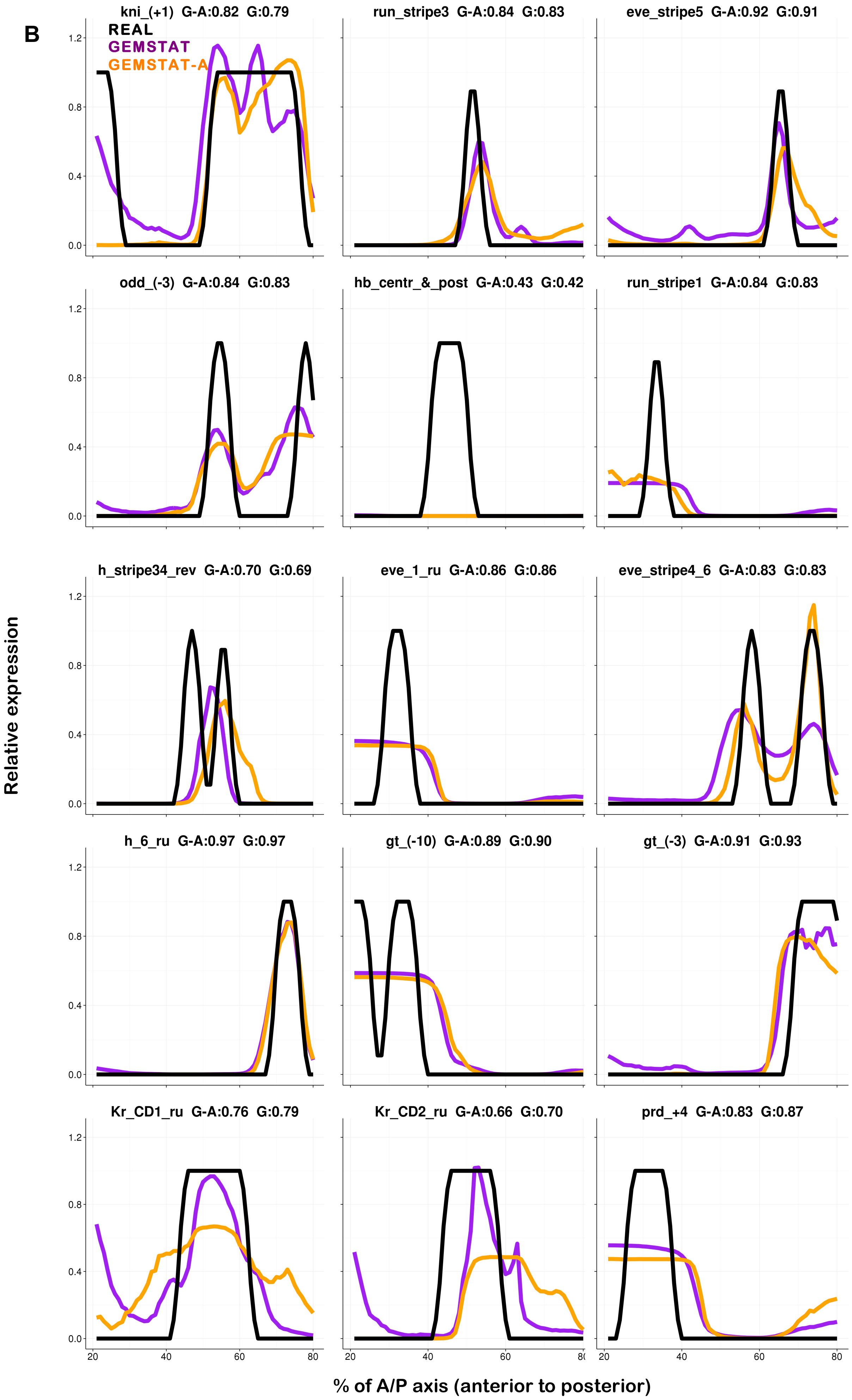


FIGURE S1 TF concentrations (y-axis) for *BCD*, *CAD*, *GT*, *HB*, *KNI*, *KR* along the A/P axis (x-axis).


A
ftz_+3 G-A:0.87 G:0.47
REAL
GEMSTAT
GEMSTAT-A
odd_(-5) G-A:0.75 G:0.45
nub_(-2) G-A:0.81 G:0.53
pdm2_(+1) G-A:0.90 G:0.64
run_stripe5 G-A:0.77 G:0.54
eve_37ext_ru G-A:0.98 G:0.79
h_15_ru G-A:0.64 G:0.46
D_(+4) G-A:0.76 G:0.60
eve_stripe2 G-A:0.79 G:0.66
kni_83_ru G-A:0.82 G:0.71
gt_(-1) G-A:0.71 G:0.60
btd_head G-A:0.89 G:0.78
hb_anterior_actv G-A:0.92 G:0.84
slp2_(-3) G-A:0.95 G:0.88
knrl_(+8) G-A:0.51 G:0.44
1.2
0.8
0.4
0.0
20
40
60
80
Relative expression
% of A/P axis (anterior to posterior)


B
REAL
GEMSTAT
GEMSTAT-A
kni_(+1) G-A:0.82 G:0.79
run_stripe3 G-A:0.84 G:0.83
eve_stripe5 G-A:0.92 G:0.91
odd_(-3) G-A:0.84 G:0.83
hb_centr_&_post G-A:0.43 G:0.42
run_stripe1 G-A:0.84 G:0.83
h_stripe34_rev G-A:0.70 G:0.69
eve_1_ru G-A:0.86 G:0.86
eve_stripe4_6 G-A:0.83 G:0.83
h_6_ru G-A:0.97 G:0.97
gt_(-10) G-A:0.89 G:0.90
gt_(-3) G-A:0.91 G:0.93
Kr_CD1_ru G-A:0.76 G:0.79
Kr_CD2_ru G-A:0.66 G:0.70
prd_+4 G-A:0.83 G:0.87
1.2
0.8
0.4
0.0
20
40
60
80
Relative expression
% of A/P axis (anterior to posterior)

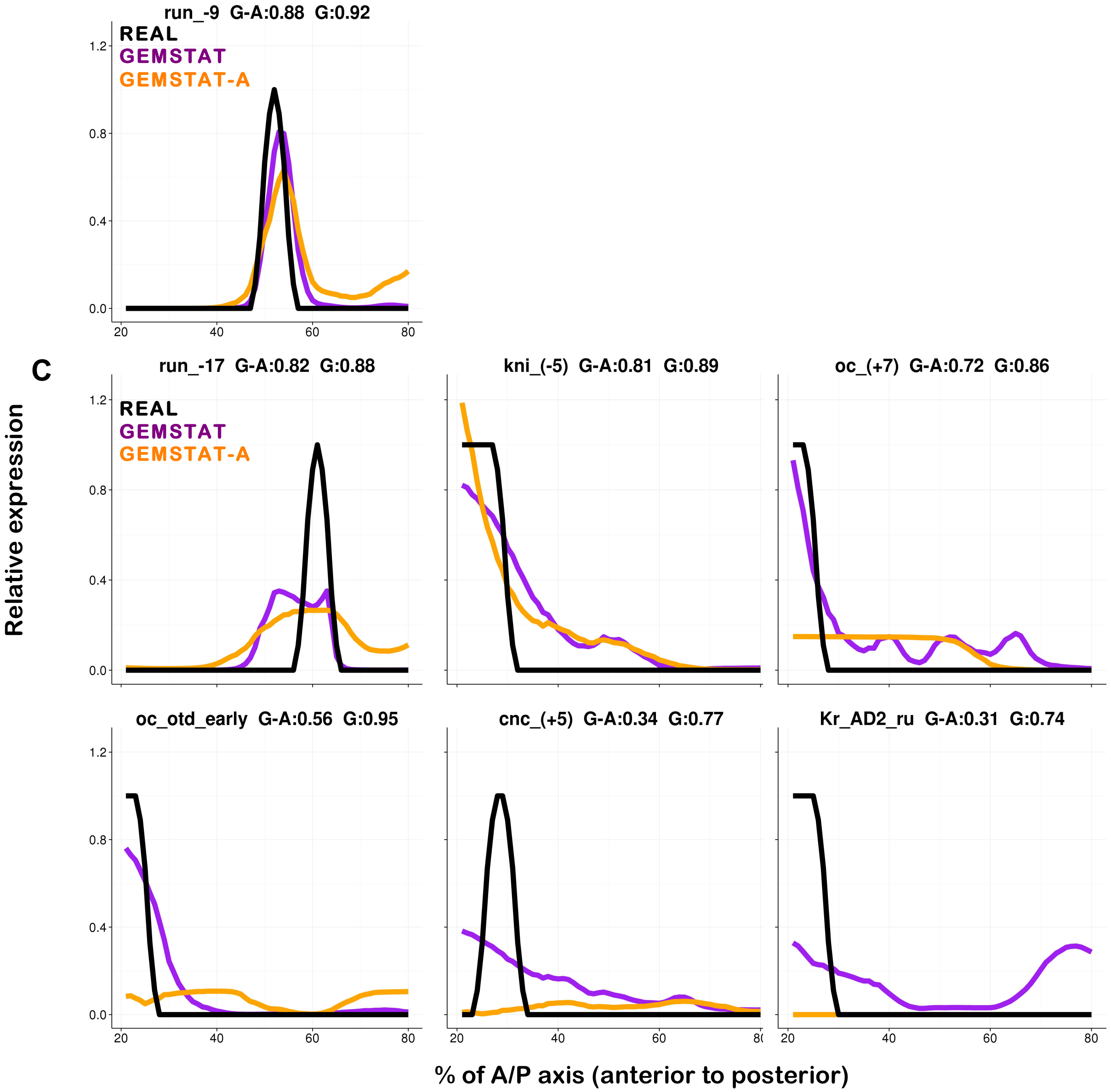


FIGURE S2 Expression predictions from GEMSTAT and GEMSTAT-A. The predicted expression profiles of GEMSTAT-A (orange lines) and GEMSTAT (purple lines) are compared to experimentally determined readouts (black lines), for 9 selected CRMs. Each expression profile is on a relative scale of 0 to 1 (y-axis), and shown for the region between 20% egg length and 80% egg length along the A/P axis of the embryo. Title in each panel is in the format of "enhancer, wPGP by GEMSTAT-A (G-A), wPGP by GEMSTAT (G)." (A) 15 enhancers with wPGP score improved by ≥ 0.05. (B) 16 enhancers with no substantial change.(C) 6 enhancers with wPGP scores worsened by ≥ 0.05. The order of enhancers is the same as in TABLE S1.

TABLE S1. Evaluations of expression predictions from GEMSTAT and GEMSTAT-A. The “goodness of fit” between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from GEMSTAT and GEMSTAT-A over all 37 enhancers are shown, and wPGP scores greater than 0.75 are colored in red.

Enhancer	GEMSTAT-A wPGP	GEMSTAT wPGP	Change ≥ 0.05	Change ≥ 0.05 and both ≥ 0.50
ftz +3	0.87	0.47	+	
odd_(-5)	0.75	0.45	+	
nub_(-2)	0.81	0.53	+	+
pdm2_(+1)	0.90	0.64	+	+
run_stripe5	0.77	0.54	+	+
eve_37ext_ru	0.98	0.79	+	+
h_15_ru	0.64	0.46	+	
D_(+4)	0.76	0.60	+	+
eve_stripe2	0.79	0.66	+	+
kni_83_ru	0.82	0.71	+	+
gt_(-1)	0.71	0.60	+	+
btd_head	0.89	0.78	+	+
hb_anterior_actv	0.92	0.84	+	+
slp2_(-3)	0.95	0.88	+	+
knrl_(+8)	0.51	0.44	+	
kni_(+1)	0.82	0.79		
run_stripe3	0.84	0.83		
eve_stripe5	0.92	0.91		
odd_(-3)	0.84	0.83		
hb_centr_&_post	0.43	0.42		
run_stripe1	0.84	0.83		
h_stripe34_rev	0.70	0.69		
eve_1_ru	0.86	0.86		
eve_stripe4_6	0.83	0.83		
h_6_ru	0.97	0.97		
gt_(-10)	0.89	0.90		
gt_(-3)	0.91	0.93		
Kr_CD1_ru	0.76	0.79		
Kr_CD2_ru	0.66	0.70		
prd_+4	0.83	0.87		
run_-9	0.88	0.92		
run_-17	0.82	0.88	-	-
kni_(-5)	0.81	0.89	-	-
oc_(+7)	0.72	0.86	-	-
oc_otd_early	0.56	0.95	-	
cnc_(+5)	0.34	0.77	-	
Kr_AD2_ru	0.31	0.74	-	-

TABLE S2. GEMSTAT-A learns stronger parameters than GEMSTAT on the same data set. The bindingWt and txpEffect parameters of each TF learned from GEMSTAT-A and GEMSTAT are shown.

TF	GEMSTAT-A bindingWt	GEMSTAT bindingWt	GEMSTAT-A txpEffect	GEMSTAT txpEffect
BCD	27.38	23.70	3.18	1.61
CAD	161.62	45.51	2.47	1.06
GT	499.98	490.17	0.01	0.07
HB	211.45	3.89	0.40	0.01
KNI	117.55	8.58	0.01	0.03
KR	264.23	253.64	0.02	0.39

TABLE S3. 10-fold cross-validation assessment. GEMSTAT and GEMSTAT-A models were tested with 10-fold cross-validation 5 times. For each 10-fold cross-validation run, the wPGP scores of GEMSTAT and GEMSTAT-A (averaged over 37 enhancers, “Avg. wPGP”) are shown.

Run #	GEMSTAT Avg. wPGP	GEMSTAT-A Avg. wPGP
1	0.676	0.748
2	0.666	0.745
3	0.685	0.736
4	0.684	0.742
5	0.685	0.737

TABLE S4. Effect of shuffling DNA accessibility data used in GEMSTAT-A. GEMSTAT-A was applied with two different types of shuffled DNA accessibility data: shuffled across whole genome and shuffled across all 37 enhancers. For each runs of shuffled DNA accessibility data, the average wPGP (“Avg. wPGP”) is shown.

Run #	Shuffling across whole genome	Shuffling across all enhancers
1	0.739	0.735
2	0.732	0.731
3	0.733	0.739

TABLE S5 Parameters used in GEMSTAT.

Parameter	Description	Number
bindingWt_i	Represents the dissociation constant of the (equilibrium) reaction between the i-th TF, TF_i and its optimal binding site when the concentration of TF_i is maximum	One per TF
q_{BTM}	A phenomenological parameter that captures the combined effect of all molecular species that act downstream of the TF recruitment step and initiate transcription (such molecular species are collectively known as the basal transcription machinery or BTM)	One global parameter
txpEffect_i	Represents the strength of TF_i's effect on the BTM	One per TF
$\omega_{i,j}$	Strength of interaction between molecules of two TFs, TF_i and TF_j (i and j may be the same), which are assumed to bind cooperatively to the DNA	One per pair of TFs (TF_i and TF_j) that are assumed to have cooperativity in DNA binding

SUPPORTING REFERENCES

1. Samee, A.H., and S. Sinha. 2013. Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data. Methods. 62: 79–90.

2. He, X., M.A.H. Samee, C. Blatti, and S. Sinha. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput. Biol. 6: e1000935.